International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST52411221

Alzheimer Disease Using Machine Learning

[1]S Dennis Emmanuel, [2]Dr G Manikandan, [3]Ms. Vilma Veronica, [4]Ms. S. Hemalatha
[1]PG Student, Kings Engineering College, Sriperumbudhur, Tamil Nadu, India
[2]Professor, Kings Engineering College, Sriperumbudhur, Tamil Nadu, India, India
[3]Assistant Professor, Kings Engineering College, Sriperumbudhur, Tamil Nadu, India
[4]Assistant Professor, Kings Engineering College, Sriperumbudhur, Tamil Nadu, India

ARTICLEINFO

Article History:

Accepted: 03 March 2024
Published: 13 March 2024

Publication Issue :
Volume 11, Issue 2
March-April-2024
Page Number :
109-113

ABSTRACT

The successful development of amyloid-based biomarkers and tests for Alzheimer's disease (AD) represents an important milestone in AD diagnosis. However, two major limitations remain. Amyloid-based diagnostic biomarkers and tests provide limited information about the disease process and they are unable to identify individuals with the disease before significant amyloid-beta accumulation in the brain develops. The objective in this study is to develop a method to identify potential blood-based non-amyloid biomarkers for early AD detection. The use of blood is attractive because it is accessible and relatively inexpensive. Our method is mainly based on machine learning (ML) techniques (support vector machines in particular) because of their ability to create multivariable models by learning patterns from complex data. Using novel feature selection and evaluation modalities we identified 5 novel panels of non-amyloid proteins with the potential to serve as biomarkers of early AD. In particular, we found that the combination of A2M, ApoE, BNP, Eot3, RAGE and SGOT may be a key biomarker profile of early disease. Disease detection models based on the identified panels achieved sensitivity (SN) > 80%, specificity (SP) > 70%, and area under receiver operating curve (AUC) of at least 0.80 at prodromal stage (with higher performance at later stages) of the disease. Existing ML models performed poorly in comparison at this stage of the disease suggesting that the underlying protein panels may not be suitable for early disease detection. Our results demonstrate the feasibility of early detection of AD using non-amyloid based biomarkers.

Index Terms—Alzheimer's disease, blood biomarker, dementia, machine learning, support vector machine.

I. INTRODUCTION

Azheimer's disease (AD) is the leading cause of dementia and poses a significant social and economic challenge. It is responsible for more than half of all cases of dementia . Over 50 million individuals currently suffer from dementia worldwide with a projected increase to 152 million by 2050. No cure for AD has been discovered, but there is intense effort to develop new clinical interventions that may slow or halt the disease. Such interventions are aimed at early (including preclinical and prodromal) stages of the disease prior to extensive cell damage, when it is thought treatment is more likely to be effective. To facilitate early diagnosis, the use of established biomarkers such as those based on amyloid-beta in cerebral spinal fluid (CSF) and molecular imaging of brain amyloid deposition using positron emission tomography (PET) is recommended. However, despite progress with the development of amyloid-based biomarkers and tests for early AD diagnosis, they have two major constraints. Amyloid-based biomarkers provide limited information about disease pathological aetiology and pathways. In addition, tests based on these biomarkers are unable to identify individuals at risk of AD prior to a significant amyloid-beta deposition in the brain.

There is a need for biomarkers that have the potential to detect biological processes that precede brain amyloid-beta accumulation (amyloid pathology) during the disease development. Such biomarkers may advance understanding of the disease, aid identification of individuals at the early disease stages and the development of new interventions. Studies suggest that AD is characterised by metabolic alterations that mayprecede amyloid pathology . Signatures of such metabolic abnormalities may therefore serve as biomarkers of earlier stages of the disease than amyloid

biomarkers. Such biomarkers may be obtained from blood since blood has rich metabolic information content. The use of blood is also attractive because blood biomarker-based test is relatively non-invasive compared to CSF and may be more cost-effective than PET imaging. A number of studies have attempted to find non-amyloid biomarkers of disease by profiling a large array of non-amyloid proteins in blood and examining their association with the disease, but this approach is difficult to apply in practice.

A promising approach is the use of machine learning (ML) techniques to find appropriate combinations of non-amyloid proteins to detect AD as no single non-amyloid protein has been shown to reliably detect the disease. ML makes it possible to fit multivariable data to a model by learning complex patterns from data. Several studies have applied ML to develop classifiers to differentiate between AD subjects and healthy controls. For example, O'Bryant *et al.* developed a model with a panel of 30 serum proteins that classified Alzheimer's disease dementia (ADD) subjects and HCs with sensitivity (SN), specificity (SP), and area under receiver operating curve (AUC) of 88%, 82%, and 0.91, respectively. Similarly, with 14 plasma proteins, a classifier model constructed by Llano *et al.* classified ADD and HC subjects with 86.5% SN, 84.2% SP and AUC of 0.85. More recently, a panel of inflammatory markers in plasma was identified that classified ADD and HC with 84% SN, 70% SP, and AUC of 0.79 using a logistic regression model. In another study, a 12-marker panel classified ADD and HC with 90% SN and 66.7% specificity, and higher performance in post-mortem confirmed AD cases. Furthermore, a study that explored the use of deep learning, random forest, and XGBoost algorithms for classification of ADD and HC achieved AUC of 0.88

with XGBoost algorithm and 0.85 with deep learning and random forest. Despite the promising results from these studies, most of the models were developed and evaluated using data from cognitively healthy controls and subjects at the later stages of the disease. The models were not evaluated in individuals at the early stages of the disease. Therefore, the panels underlying such models may not be suitable as biomarker signatures of early AD.

In this study, the main objective is to develop a ML-based method (support vector machines (SVM) in particular – see later) to identify blood biomarkers of early AD based on non-amyloid proteins with the potential to identify the disease prior to accumulation of amyloid-beta in the brains.

We also assess the potential of existing ML-based methods to achieve early disease detection.

The rest of this paper is structured as follows. The materials and methods are described in Sections II and III. The results are presented in Section IV, and the discussion and conclusions are provided in Sections V and VI.

II. MATERIALS

Blood proteomic data used in this study were obtained from the Alzheimer's disease neuroimaging initiative (ADNI) portal

TABLE I. DEMOGRAPHIC INFORMATION OF SUBJECTS IN STUDY DATA

Clinical groups	Sample size	Ave. age in years (SD)	Ave. years of education (SD)	% Female
HC	58(54)	75(6)	16(2.8)	48(50)
MCI	136	75(7)	16(3.0)	45
ADD	108	75(8)	15(3.2)	46

(http://adni.loni.ucla.edu). The quality-controlled data consist of 146 plasma proteins derived from 58 and 54 healthy controls (HCs) at baseline and 12 months later respectively, 136 individuals with mild cognitive impairment due to AD (MCI) at 12 months from baseline, and 108 Alzheimer's dementia (ADD) patients at baseline. The MCI subjects were later diagnosed with AD dementia within about 10-year follow-up. A list of the 146 proteins are shown in the supplementary material. Mild dementia was diagnosed according to NINCDS-ADRDA criteria for probable ADD. A detailed description of the protocol may be found on the ADNI database. The demographic information of the subjects is shown in Table I. The subjects were age matched, over 70 years old and had about 16 years of education on average.

III. PROPOSED SYSTEM

A. Data pre-processing

To make optimal use of available data while minimizing susceptibility of our approach to overfitting problems, the pre-processed data were partitioned into two non-overlapping datasets; Datasets 1 and 2. Dataset 1 consists of baseline data from the ADDs and HCs. All existing methods evaluated in this study except were originally developed based on Dataset 1. In our approach, Dataset 1 was used to conduct a robust feature preselection (a key aspect in ML) and model development.

The resulting models were further evaluated with Dataset 2. Dataset 2 consists of month-12 data from MCIs and HCs. It was used to assess the performance of the developed models (trained on the entirety of Dataset 1) for MCI vs. HC classification. Models were trained with only Dataset 1 during model development using the entirety of it or its subsamples (in the case of cross-validation which is subsequently described).

B. Replication and evaluation of existing methods

We replicated the ML models reported in previous studies for classification of ADD and HC subjects (Dataset 1) using 10- fold cross-validation with the

average performance of the models taken after 10 repetitions. In 10-fold cross-validation, the dataset D is randomly split into 10 mutually exclusive subsets (the folds) D1, D2, ..., D10 of approximately equal size. The classifier is trained and tested 10 times; each time $t \in \{1, 2, \dots, 10\}$, it is trained on D\Dt and tested on Dt. The cross-validation estimate of the classifier performance is the overall performance over all the folds. Repeated cross-validation was implemented to ensure a robust estimation of performance. The ability of the models to classify MCI and HC was then tested with Dataset 2 to assess their potential and hence the underlying protein panels to detect early AD.

C. Novel panel identification and model development

The methodological framework that we used to identify novel blood protein panels and to develop the new ML models for early detection of AD. The framework is described in detail in the following subsections. Briefly, the framework consists of three major procedures which include feature subset preselection, protein panel formation, and ML-based model development and evaluation. A feature subset preselection process was performed to identify protein subsets that may have strong discriminatory power between disease subjects (ADD) and HCs. A brute force search was applied to the preselected feature subset to form several protein panels. Each of the panels was then used to develop and cross-validate SVM classifiers of different kernels (K) using Dataset 1. Data from ADD subjects were used in these initial procedures on the basis that dementia subjects are more likely to exhibit the metabolic alterations that are associated with the disease. The most stable kernel and candidate panels (promising models) trained on Dataset 1 were further evaluated for classification of individuals with MCI and HCs using Dataset 2. The promising models with best performance at this stage were selected as final. The protein panels that underlie the selected models are reported as potential blood-based non-amyloid biomarker signature of early disease.

D. Implementation and performance evaluation

Feature selection using CFS as discussed earlier was conducted with attribute selection toolbox in Weka software package. All classification tasks were conducted with MATLAB and Weka software packages. In evaluating the models from previous studies, we used Weka where previous studies had used it for model development. Training of ML models and validation of performance for ADD vs. HC discrimination was based on 10-fold cross-validation scheme repeated 10 times. The data (Dataset 1) were randomly re-partitioned after each run to ensure that data subsets used for training and validation varied from the ones used in the preceding run. This way, a more robust average performance is obtained. Classification performance metrics of primary consideration were measures of SN and SP in accordance with international recommendations for clinically usable AD biomarkers. A performance threshold of 70% for SN and SP was adopted in the model development task. This is on the grounds that the diagnostic accuracy of human experts reaches 77% with sensitivity and specificity reaching 81% and 70%, respectively. Moreover, sensitivity and specificity greater than 80% is the target performance for ideal AD biomarkers. No class imbalance handling procedure was applied to the training dataset (Dataset 1) in model development as minority to majority class distribution was 35:65% which is acceptable in ML-based classification problems.

IV. Conclusion

We have developed potential models and identified five novel candidate non-amyloid biomarker panels for early detection of AD utilizing a new approach. The developed models based on these panels classified prodromal AD as well as AD dementia and normal controls with sensitivity above 80%, specificity higher than 70%, and AUC of at least 0.80. A combination of A2M, ApoE, BNP, Eot3, RAGE and SGOT were

identified as key protein profiles with significant contribution to the classifications performance. The results suggest that it may be feasible to detect early AD using a profile of non-amyloid proteins that identify the metabolic processes that accompany or precede the disease. It may be therefore possible to detect the disease with the proteins before amyloid pathology (the earliest signature current diagnostic biomarkers can detect) develops since they are not amyloid-based. This may aid identification of individuals at the earliest stages of AD who may benefit from early interventions. Furthermore, new insights about the disease may be gained from understanding the interactions between the proteins in disease subjects. Such enhanced understanding may contribute to the improvement of interventions in clinical trials.

V. REFERENCES

[1]. A. Association, "2018 Alzheimer's disease facts and figures," Alzheimer's & Dementia, vol. 14, no. 3, pp. 367-429, 2018.

[2]. M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World Alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future," 2016.

[3]. B. Dubois et al., "Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria," Alzheimer's & Dementia, vol. 12, no. 3, pp. 292-323, 2016.

[4]. M. S. Albert et al., "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," Alzheimer's & Dementia, vol. 7, no. 3, pp. 270-279, 2011.

[5]. G. M. McKhann et al., "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," Alzheimer's & Dementia, vol. 7, no. 3, pp. 263-269, 2011.

[6]. R. A. Sperling et al., "Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," Alzheimer's & Dementia, vol. 7, no. 3, pp. 280-292, 2011.

[7]. G. P. Morris, I. A. Clark, and B. Vissel, "Questions concerning the role of amyloid-β in the definition, aetiology and diagnosis of Alzheimer's disease," Acta neuropathologica, vol. 136, no. 5, pp. 663-689, 2018.

[8]. K. H. Tse and K. Herrup, "Re-imagining Alzheimer's disease–the diminishing importance of amyloid and a glimpse of what lies ahead," Journal of Neurochemistry, vol. 143, no. 4, pp. 432-444, 2017.

[9]. F. Zhang, J. Wei, X. Li, C. Ma, and Y. Gao, "Early Candidate Urine Biomarkers for Detecting Alzheimer's Disease Before Amyloid-β Plaque Deposition in an APP (swe)/PSEN1 dE9 Transgenic Mouse Model," Journal of Alzheimer's Disease, no. Preprint, pp. 1-25, 2018.

[10]. F. Kametani and M. Hasegawa, "Reconsideration of amyloid hypothesis and tau hypothesis in Alzheimer's disease," Frontiers in neuroscience, vol. 12, p. 25, 2018.

Cite this article as :

S Dennis Emmanuel, Dr G Manikandan, Ms. Vilma Veronica, Ms. S. Hemalatha, "Alzheimer Disease Using Machine Learning", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 11 Issue 2, pp. 109-113, March-April 2024. Available at doi : https://doi.org/10.32628/IJSRST52411221

Journal URL : https://ijsrst.com/IJSRST52411221